IE 498: Online Learning and Decision Making, Fall 2019 10-22-2019

Lecture-17: Model Based Reinforcement Learning

Lecturer: Yuan Zhou *Scribe: Xiaobo Dong, Yufei Ruan*

1 Motivation

We consider the reinforcement learning (RL) problem of an agent interacting with an environment in order to maximize its cumulative rewards through time. We model the environment as a Markov decision process (MDP) whose transition dynamics are unknown from the agent. As the agent interacts with the environment it observes the states, actions and rewards generated by the system dynamics. This leads to a fundamental trade off: should the agent explore poorly-understood states and actions to gain information and improve future performance, or exploit its knowledge to optimize short-run rewards. We show that an optimistic modification to value iteration achieves a regret bound of $\mathcal{O}(H^{1.5}|S|\sqrt{|A|T})$ (where O ignores logarithmic factors).

2 Assumption

In this algorithm, we have the following assumptions:

1. State S and action A are finite sets with cardinality $|S|$,$|A|$ respectively.
2. Assume The reward $R(x, a)$ is immediate deterministic and known. $R(x, a) \in [0, 1]$

3 Notation

- S: State Space
- A: Action Space
- H: Time-horizon length
- π: $S \times [H] \to A$
- $V_h^\pi(x)$: The value function starting from x in at step h with policy π.
- $V_h^*(x)$: The optimal value function starting from x in at step h. $V_h^*(x) = \max_\pi V_h^\pi(x)$
- $Q_h^\pi(x, a)$: The state-action value function starting from x in at step h with policy π.
- $Q_h^*(x, a)$: The optimal state-action value function starting from x with action a in at step h.
- $P_h^\pi(y|x) = P(y|x, \pi(x, h))$: Unknown environment dynamics.
- $r_h^\pi(x) = R(x, \pi(x, h)$: The deterministic immediate reward.

- $R_K = \sum_{k=1}^{K} V_1^*(x_{k,1}) - V_1^{\pi_k}(x_{k,1})$: The regret up to time K.
- $x_{k,h}$: State $x \in S$ at step h in episode k.
- $V_{k,h}$: Estimated V_h^* at step h in episode k.
- $Q_{k,h}$: Estimated Q_h^* at step h in episode k.
- $N_{k,h}(x,a)$: Number of (x,a) pairs up to k.
- $N_{k,h}(x,a)$: Number of (x,a,y) tuples up to k.
- $\hat{P}_{k,h}(y|x,a)$: Estimated $P(y|x_{k,h},a)$ at step h in episode k.
- $PV(x,a) = \sum_{y \in S} P(y|x,a)V(y) =< P(y|x,a), V(y) >$
- $K = \{(x,a) \in S \times A, N_{k,h}(x,a) > 0\}$
- C: Some constant factor that we ignore during the calculation.

Note though the unknown transition matrix P is stationary in finite MDP, the estimate transition matrix $\hat{P}$ has to be a function of h because the policy function has to be a function of h.

4 Algorithm

The Upper Confident Bound Value Iteration algorithm(UCBVI) described in Algorithm 1 calls UCBQ described in Algorithm 2. UCBQ computes Q-values by value iteration using an empirical Bellman operator to which is added a confidence bonus $b_{k,h}$. Moreover, the bonus function $b_{k,h}$ is given by Hoeffding's concentration inequality, which will be shown in next section.

Algorithm 1: UCBVI

Initialize data buffer $H = \varnothing$;
for *epoch* $k = 1, 2, ...K$ **do**
  $Q_{k,h} =$ **UCBQ**(H) **for** *step h = 1,2,...,H* **do**
    - Take action $a_{k,h} = \underset{a \in A}{argmax}\ Q_{k,h}(x_{k,h}, a)$
    - Sample $x_{k,h+1}$ from the dynamics of the Environment.
    - $H = H \cup (x_{k,h}, a_{k,h}, x_{k,h+1})$
  end
end

Algorithm 2: UCBQ

Data: Data H
Result: Q-value $Q_{k,h}$

Initialize $V_{k,H+1}(x) = 0$ for all $x \in S$;
Estimate $\hat{P}_{k,h}(y|x,a) = \begin{cases} \frac{N_{k,h}(x,a,y)}{N_{k,h}(x,a)}, & \text{for } (x,a) \in K \\ 0, & \text{otherwise} \end{cases}$;
Calculate "Bonus" $b_{k,h}(x,a) = \begin{cases} CH\sqrt{\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}, & \text{for } (x,a) \in K \\ H, & \text{otherwise} \end{cases}$;
Estimate $Q_{k,h}$ and $V_{k,h}$;
for *h = H,H-1,...,1* **do**
  for $(x,a) \in S \times A$ **do**
    - $Q_{k,h} = \begin{cases} R(x,a) + (\hat{P}_{k,h}V_{k,h+1})(x.a) + b_{k,h}(x,a), & \text{for } (x,a) \in K \\ b_{k,h}(x,a), & \text{otherwise} \end{cases}$
    - $V_{k,h} = \max_{a \in A} Q_{k,h}(x,a)$
  end
end
return $Q_{k,h}$

5 Proof of the Upper Bound of Regret

The regret of UCBVI is given by $R_K = \sum_{k=1}^{K} V_{k,1}^*(x) - V_{k,1}^{\pi_k}(x)$, and We define $\delta_{k,h} = (V_{k,h} - V_h^{\pi_k})(x)$. Note that the $\delta_{k,h}$ is defined as the difference between the estimated value function and value function given by the policy π_k. The agenda of the proof is as following: Fist, given by claim 1, we will get the upper bound of the regret R_K by $R_K \leq \mathbb{E}\sum_{k=1}^{K} \delta_{k,1} + \delta T$. Second, we will get the upper bound of $\delta_{k,h}$ by Bernstein inequality. In the end, we will wrap everything together to get the upper bound of the regret.

Claim 1. *Define event* $\mathcal{G} = \left\{ \begin{cases} Q_{k,h}(x,a) \geq Q_h^*(x,a) \\ V_{k,h}(x) \geq V_h^*(x) \end{cases} \quad \forall k,h,x,a \right\}$

$$P(\mathcal{G}) \geq 1 - \delta$$

Proof. Prove by backward induction

- Base case: For $\hat{h} = H$

$$Q_{k,H}(x,a) = \begin{cases} R(x,a) + b_{k,H}(x,a), & \text{for } (x,a) \in K \\ H, & \text{otherwise} \end{cases}$$
$$Q_H^*(x,a) = R(x,a)$$
$$\Rightarrow Q_{k,H}(x,a) \geq Q_H^*(x,a) \qquad \forall (x,a) \in S \times A$$

  And thus,

$$V_{k,H}(x) = \max_a Q_{k,H}(x,a) \geq V_H^*(x) = \max_a Q_H^*(x,a)$$

- Assume for $\hat{h} = H, H-1, ..., h+1$ hold

- Now to show $\hat{h} = h$ holds

$$\begin{aligned} Q_{k,h}(x,a) - Q_h^*(x,a) &= \left(\hat{P}_{k,h}V_{k,h+1}\right)(x,a) + b_{k,h}(x,a) - \left(PV_{h+1}^*\right)(x,a) \\ &= \left(\hat{P}_{k,h}V_{k,h+1}\right)(x,a) + b_{k,h}(x,a) - \left(PV_{h+1}^*\right)(x,a) + \left(\hat{P}_{k,h}V_{h+1}^*\right)(x,a) - \left(\hat{P}_{k,h}V_{h+1}^*\right)(x,a) \\ &= \left(\hat{P}_{k,h}(V_{k,h+1} - V_{h+1}^*)\right)(x,a) + b_{k,h}(x,a) + \left((\hat{P}_{k,h} - P)V_{h+1}^*\right)(x,a) \end{aligned}$$

By induction we know $\left(\hat{P}_{k,h}(V_{k,h+1} - V_{h+1}^*)\right)(x,a) \geq 0$. Now is to show that the remaining part $b_{k,h}(x,a) + \left((\hat{P}_{k,h} - P)V_{h+1}^*\right)(x,a)$ is non-negative with a high probability.

Define the following events

$$\mathcal{E} = \left\{ \left| \left((\hat{P}_{k,h} - P)V_{h+1}^*\right)(x,a) \right| \leq b_{k,h}(x,a) \quad \forall k,h,x,a \right\}$$

$$\mathcal{E}_{k,h,x,a} = \left\{ \left| \left((\hat{P}_{k,h} - P)V_{h+1}^*\right)(x,a) \right| \leq b_{k,h}(x,a) \right\}$$

Let $\mathcal{E}^C$ and $\mathcal{E}_{k,h,x,a}^C$ denote the corresponding complementary of the events.

Theorem 1: Union Bound

For A_1,A_2,...,A_n, we have

$$\mathbb{P}\left(\cup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \mathbb{P}\left(A_i\right)$$

Given by the Union Bound, we know

$$\mathbb{P}(\mathcal{E}^C) \leq \sum_{k,h,x,a} \mathbb{P}(\mathcal{E}_{k,h,x,a}^C) = |S||A|T \cdot \mathbb{P}(\mathcal{E}_{k,h,x,a}^C)$$

Theorem 2: Hoeffding Inequality

Let Z_1,Z_2,...,Z_n be independent bounded random variables with $Z_i \in [a,b]$ for all i. Then we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (Z_i - E[Z_i])\right| \geq t\right) \leq 2e^{-\frac{2nt^2}{(b-a)^2}}$$

Moreover,given by the Hoeffding Bound, we have

$$\begin{aligned}
\mathbb{P}(\mathcal{E}^C_{k,h,x,a}) &= \mathbb{P}\bigg(\bigg|\Big((\hat{P}_{k,h} - P)V^*_{h+1}\Big)(x,a)\bigg| \geq b_{k,h}(x,a)\bigg) \\
&= \mathbb{P}\bigg(\bigg|\sum_{y\in S}\Big(\hat{P}_{k,h}(y|x,a) - P(y|x,a)\Big)V^*_{h+1}(y)\bigg| \geq b_{k,h}(x,a)\bigg) \\
&= \mathbb{P}\bigg(\bigg|\sum_{y\in S}\hat{P}_{k,h}(y|x,a)V^*_{h+1}(y) - \sum_{y\in S}P(y|x,a)V^*_{h+1}(y)\bigg| \geq b_{k,h}(x,a)\bigg) \\
&\text{Let } Z = \sum_{y\in S}\hat{P}_{k,h}(y|x,a)V^*_{h+1}(y), \text{ and then } E[Z] = \sum_{y\in S}P(y|x,a)V^*_{h+1}(y) \\
&\leq 2e^{-2\frac{b^2_{k,h}(x,a)}{H^2}\cdot N_{k,h}(x,a)}
\end{aligned}$$

Thus, we have

$$\mathbb{P}(\mathcal{E}^C) \leq |S||A|T \cdot \mathbb{P}(\mathcal{E}^C_{k,h,x,a}) = |S||A|T \cdot 2e^{-2\frac{b^2_{k,h}(x,a)}{H^2}\cdot N_{k,h}(x,a)}$$

If you solve $|S||A|T \cdot 2e^{-2\frac{b^2_{k,h}(x,a)}{H^2}\cdot N_{k,h}(x,a)} = \delta$, then we can find $b_{k,h} = CH \cdot \sqrt{\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}$ Thus,$\mathbb{P}(\mathcal{E}) \geq 1-\delta$. Therefore, $\hat{h} = h$ holds

Then by the backward induction, we have the result

$$\mathbb{P}(\mathcal{G}) \geq 1 - \delta$$

□

Claim 2.

$$R_K = \mathbb{E}\sum_{k=1}^{K}(V^*_1 - V^{\pi_k}_1)(x) \leq \mathbb{E}\sum_{k=1}^{K}(V_{k,1} - V^{\pi_k}_1)(x) + \delta T = \mathbb{E}\sum_{k=1}^{K}\delta_{k,1} + \delta T$$

where $\delta_{k,h} \triangleq (V_{k,h} - V^{\pi_k}_h)(x)$

Proof. Given by Claim 1, we know event $\mathcal{E}$ holds with probability bigger than $1-\delta$.

$$\begin{aligned}
\mathbb{E}[R_K] &= \mathbb{E}[R_K|\mathcal{E}] \times (1-\delta) + \mathbb{E}[R_K|\mathcal{E}^C] \times \delta \\
&\leq \mathbb{E}\sum_{k=1}^{K}(V_{k,1} - V^{\pi_k}_1)(x) + \delta T
\end{aligned}$$

where R_K bounded by $T = H \times K$ is used in the second term. □

Claim 3. $\mathbb{E}[\delta_{k,h}] = \mathbb{E}\left[\Big((\hat{P}_{k,h} - P)V_{k,h+1}\Big)(x,a) + b_{k,h}(x,a)\right] + \mathbb{E}[\delta_{k,h+1}]$

Proof. By definition, we have $\delta_{k,h} \triangleq \Big(V_{k,h} - V_h^{\pi_k}\Big)(x) = \Big(Q_{k,h} - Q^{\pi_k}\Big)(x,a)$

$$\begin{aligned}
\delta_{k,h} &= \Big(Q_{k,h} - Q^{\pi_k}\Big)(x,a) \\
&= Q_{k,h}(x,a) - Q^{\pi_k}(x,a) \\
&= \hat{P}_{k,h}V_{k,h+1}(x,a) + b_{k,h}(x,a) - PV_{k,h+1}^{\pi_k}(x) \\
&= \hat{P}_{k,h}V_{k,h+1}(x,a) + b_{k,h}(x,a) - PV_{k,h+1}^{\pi_k}(x) - PV_{k,h+1}(x,a) + PV_{k,h+1}(x,a) \\
&= \Big((\hat{P}_{k,h} - P)V_{k,h+1}\Big)(x,a) + b_{k,h}(x,a) + \Big(PV_{k,h+1}(x,a) - PV_{k,h+1}^{\pi_k}(x))\Big)
\end{aligned}$$

Moreover, $\mathbb{E}[\delta_{k,h+1}] = \mathbb{E}\left[V_{k,h+1}(x) - V_{h+1}^{\pi_k}(x)\right] = \mathbb{E}\left[PV_{k,h+1}(x,a) - PV_{k,h+1}^{\pi_k}(x)\right]$. The last equality comes the fact the convex combination preserves the expectation. Thus, we have

$$\mathbb{E}[\delta_{k,h}] = \mathbb{E}\left[\Big((\hat{P}_{k,h} - P)V_{k,h+1}\Big)(x,a) + b_{k,h}(x,a)\right] + \mathbb{E}[\delta_{k,h+1}]$$

□

Claim 4. $\mathbb{E}\left[\Big((\hat{P}_{k,h} - P)V_{k,h+1}\Big)(x,a)\right] \leq CH\sqrt{\frac{|S|}{N_{k,h}(x,a)}\ln(\frac{|S||A|T}{\delta})} + \frac{CH|S|}{N_{k,h}(x,a)}\ln(\frac{|S||A|T}{\delta}) + \delta H$

Proof. First of all,

$$\mathbb{E}\left[\Big((\hat{P}_{k,h} - P)V_{k,h+1}\Big)(x,a)\right] \leq \mathbb{E}\left[\left|\Big((\hat{P}_{k,h} - P)V_{k,h+1}\Big)(x,a)\right|\right]$$

Then we have

$$\begin{aligned}
\left|\Big((\hat{P}_{k,h} - P)V_{k,h+1}\Big)(x,a)\right| &= \left|\sum_{y\in S}\hat{P}_{k,h}(y|x,a)V_{k,h+1}(y) - P(y|x,a)V_{k,h+1}(y)\right| &(1)\\
&\leq \sum_{y\in S}\left|\hat{P}_{k,h}(y|x,a)V_{k,h+1}(y) - P(y|x,a)V_{k,h+1}(y)\right| &(2)\\
&\leq H\sum_{y\in S}\left|\hat{P}_{k,h}(y|x_{k,h},a) - P(y|x_{k,h},a)\right| &(3)
\end{aligned}$$

Theorem 3: Bernstein inequality

Let Z_1,Z_2,...,Z_n be independent bounded random variables with $|Z_i| \leq M$ for all i. Then we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - E[Z_i])\right| \geq t\right) \leq 2e^{-\frac{\frac{1}{2}(nt)^2}{\sum_{i=1}^{n}E[(Z_i - E[Z_i])^2] + \frac{1}{3}Mnt}}$$

When Z_i are i.i.d. random variable, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - E[Z_i])\right| \geq t\right) \leq 2e^{-\frac{\frac{1}{2}nt^2}{Var(Z_i) + \frac{1}{3}Mt}}$$

When $Z_i \in [0,1]$, we have $E[Z_i] \geq E[Z_i^2] \geq Var(Z_i)$.Thus, we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - E[Z_i])\right| \geq t\right) \leq 2e^{-\frac{\frac{1}{2}nt^2}{E(Z_i) + \frac{1}{3}t}}$$

Now bound $\Big|\hat{P}_{k,h}(y|x_{k,h},a) - P(y|x_{k,h},a)\Big|$ this term with Bernstein Inequality.

$$\mathbb{P}\Big(\Big|\hat{P}_{k,h}(y|x_{k,h},a) - P(y|x_{k,h},a)\Big| \geq t\Big) \leq 2e^{-\frac{\frac{1}{2}N_{k,h}(x,a)t^2}{\mu+\frac{1}{3}t}}$$

Note the randomness comes from $\hat{P}_{k,h}(y|x_{k,h},a)$ and $E[\hat{P}_{k,h}(y|x_{k,h},a)] = P(y|x_{k,h},a) = \mu$. Now let $\frac{\delta}{|S||A|T} = e^{-\frac{\frac{1}{2}N_{k,h}(x,a)t^2}{\mu+\frac{1}{3}t}}$, and we can solve t. We get

$$t = C_1\frac{\ln(\frac{\|S||A|T}{\delta})}{N_{k,h}(x,a)} + \sqrt{C_2\frac{\ln(\frac{|S||A|T}{\delta})\mu}{N_{k,h}(x,a)} + C_3\Big(\frac{\ln(\frac{|S||A|T}{\delta})\mu}{N_{k,h}(x,a)}\Big)^2} \approx C\Big(\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)} + \sqrt{\frac{\ln(\frac{|S||A|T}{\delta})\mu}{N_{k,h}(x,a)}}\Big)$$

Note that since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we can choose C large enough such that the approximate sign "$\approx$" becomes "$\leq$". For fixed δ, the increase in t does not change the inequality sign.

Now define the following events

$$\mathcal{F}_{k,h,s,a} \triangleq \{\Big|\hat{P}_{k,h}(y|x,a) - P(y|x,a)\Big| \leq C\Big(\sqrt{\frac{\ln(\frac{|S||A|T}{\delta})P(y|s,a)}{N_{k,h}(x,a)}} + \frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\Big)\}$$

$$\mathcal{F} \triangleq \{\forall k,h,s,a\Big|\hat{P}_{k,h}(y|x,a) - P(y|x,a)\Big| \leq C\Big(\sqrt{\frac{\ln(\frac{|S||A|T}{\delta})P(y|s,a)}{N_{k,h}(x,a)}} + \frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\Big)\}$$

We know that

$$\mathbb{P}(\mathcal{F}^C_{k,h,s,a}) \leq \frac{\delta}{|S||A|T}$$

Then given by the union bound

$$\mathbb{P}(\mathcal{F}^C) = \mathbb{P}(\cup_{k,h,s,a}\mathcal{F}^C_{k,h,s,a}) \leq |S||A|T\mathbb{P}(F^C_{k,h,s,a}) = |S||A|T \times \frac{\delta}{|S||A|T} = \delta$$

Thus, we have

$$\mathbb{P}(\mathcal{F}) \geq 1 - \delta$$

Now back to equation (3), we have

$$H\sum_{y\in S}\Big|\hat{P}_{k,h}(y|x,a) - P(y|x,a)\Big| \leq H\sum_{y\in S}C\Big(\sqrt{\frac{\ln(\frac{|S||A|T}{\delta})P(y|s,a)}{N_{k,h}(x,a)}} + \frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\Big) + H\delta \tag{4}$$

Note when event $\mathcal{F}$ does not hold, $\sum_{y\in S}\Big|\hat{P}_{k,h}(y|x,a) - P(y|x,a)\Big| \leq 1$.

Theorem 4: Cauchy–Schwarz inequality

For vectors u and v, we have

$$\Big|\sum_i^n u_i v_i\Big| \leq ||u|| \cdot ||v||$$

Now use Cauchy, we have the following trick

$$\sum_{y\in S}\sqrt{P(y|x,a)} \leq \sqrt{\sum_{y\in S}P(y|x,a)\sum_{y\in S}1} = \sqrt{|S|}$$

Then, for inequality (4), we have

$$H\sum_{y\in S}\left|\hat{P}_{k,h}(y|x,a)-P(y|x,a)\right| \le H\sum_{y\in S}C\Big(\sqrt{\frac{\ln(\frac{|S||A|T}{\delta})P(y|s,a)}{N_{k,h}(x,a)}}+\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\Big)+H\delta \tag{5}$$

$$\le CH\sqrt{\frac{\ln(\frac{|S||A|T}{\delta})|S|}{N_{k,h}(x,a)}}+CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}+H\delta \tag{6}$$

Thus, we have

$$\mathbb{E}\left[\Big((\hat{P}_{k,h}-P)V_{k,h+1}\Big)(x,a)\right] \le \mathbb{E}\left[CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}+CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right]+\delta H$$

□

Using our Claim 3 and Claim 4, we can bound $\mathbb{E}[\delta_{k,h}]$ with the following Claim.

Claim 5.

$$\mathbb{E}[\delta_{k,1}] \le \mathbb{E}\sum_{h=1}^{H}\Big(CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}+CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\Big)+\delta H^2$$

Proof. Given by Claim 3, we have

$$\mathbb{E}[\delta_{k,h}]-\mathbb{E}[\delta_{k,h+1}] = \mathbb{E}\left[\Big((\hat{P}_{k,h}-P)V_{k,h+1}\Big)(x,a)+b_{k,h}(x,a)\right]$$

We know

$$\begin{aligned}
&\mathbb{E}\left[\Big((\hat{P}_{k,h}-P)V_{k,h+1}\Big)(x,a)+b_{k,h}(x,a)\right]\\
\le&\mathbb{E}\left[CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}+CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}+b_{k,h}(x,a)\right]+\delta H\\
\le&\mathbb{E}\left[CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}+CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}+CH\sqrt{\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}\right]+\delta H\\
\approx&\mathbb{E}\left[CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}}+CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right]+\delta H
\end{aligned}$$

Note that the last approximate equality comes from the fact the bonus function has the same order as the first term. With appropriate select C, the equality can hold.

We Telescoping Sum, we have

$$\begin{aligned}
\mathbb{E}[\delta_{k,1}] - \mathbb{E}[\delta_{k,2}] &\leq \mathbb{E}\left[CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}} + CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right] + \delta H \\
\mathbb{E}[\delta_{k,2}] - \mathbb{E}[\delta_{k,3}] &\leq \mathbb{E}\left[CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}} + CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right] + \delta H \\
&\vdots \\
\mathbb{E}[\delta_{k,H}] - \mathbb{E}[\delta_{k,H+1}] &\leq \mathbb{E}\left[CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}} + CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right] + \delta H
\end{aligned}$$

where $\mathbb{E}[\delta_{k,H+1}] = 0$. We have

$$\mathbb{E}[\delta_{k,1}] \leq \mathbb{E}\sum_{h=1}^{H}\left(CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}} + CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right) + \delta H^2$$

□

Claim 6.

$$R_K \leq \delta T^2 + C|S|^2|A|H^2\log^2(\frac{|S||A|T}{\delta}) + C|S|H^{1.5}\sqrt{|A|T\log(\frac{|S||A|T}{\delta})}$$

Proof. We know

$$\begin{aligned}
R_K &\leq \mathbb{E}\sum_{k=1}^{K}\delta_{k,1} + \delta T \\
&\leq \mathbb{E}\sum_{k=1}^{K}\left[\sum_{h=1}^{H}\left(CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}} + CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right) + \delta H^2\right] + \delta T \\
&\leq \mathbb{E}\sum_{k=1}^{K}\sum_{h=1}^{H}\left(CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}} + CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{N_{k,h}(x,a)}\right) + \delta T^2 \\
&= \mathbb{E}\sum_{x,a,h}\sum_{n=1}^{N_{k,h}(x,a)}\left(CH\sqrt{|S|\frac{\ln(\frac{|S||A|T}{\delta})}{n}} + CH|S|\frac{\ln(\frac{|S||A|T}{\delta})}{n}\right) + \delta T^2
\end{aligned}$$

Using the following three facts, we can bound the term above.

Fact 1.

$$\sum_{k=1}^{n}\sqrt{\frac{1}{k}} < 2\sqrt{n}$$

Proof.

$$\frac{1}{\sqrt{k}} - 2(\sqrt{k} - \sqrt{k-1})) = \frac{1}{\sqrt{k}} - \frac{2}{(\sqrt{k} + \sqrt{k-1})} < \frac{1}{\sqrt{k}} - \frac{2}{(\sqrt{k} + \sqrt{k})} = 0$$

Thus, we have

$$\frac{1}{\sqrt{k}} < 2(\sqrt{k} - \sqrt{k-1})$$

Therefore, we have

$$\sum_{k=1}^{n} \sqrt{\frac{1}{k}} < 2\sqrt{n}$$

□

Fact 2. (It is trial that summation is smaller than integral.)

$$\sum_{k=1}^{n} \frac{1}{k} \approx \ln(n)$$

Fact 3.(Cauchy-Schwarz inequality)

$$\sum_{s,a,h} \sqrt{N_{k,h}} \leq \sqrt{\sum_{s,a,h} N_{k,h}} \sqrt{\sum_{s,a,h} 1} = \sqrt{T}\sqrt{|S||A|H}$$

Then

$$\begin{aligned}
&\mathbb{E} \sum_{x,a,h} \sum_{n=1}^{N_{k,h}(x,a)} \left(CH\sqrt{|S| \frac{\ln(\frac{|S||A|T}{\delta})}{n}} + CH|S| \frac{\ln(\frac{|S||A|T}{\delta})}{n} \right) + \delta T^2 \\
\leq &\mathbb{E} \Big[\sum_{x,a,h} \left(CH\sqrt{|S| N_{k,h}(x,a) \ln(\frac{|S||A|T}{\delta})} + CH|S| \ln^2(\frac{|S||A|T}{\delta}) \right) \Big] + \delta T^2 \\
= &\mathbb{E} \Big[\sum_{x,a,h} CH\sqrt{|S| N_{k,h}(x,a) \ln(\frac{|S||A|T}{\delta})} \Big] + CH^2|A||S|^2 \ln^2(\frac{|S||A|T}{\delta}) + \delta T^2 \\
\leq &CH\sqrt{T|S|^2|A|H \ln(\frac{|S||A|T}{\delta})} + CH^2|A||S|^2 \ln^2(\frac{|S||A|T}{\delta}) + \delta T^2 \\
= &CH^{1.5}|S|\sqrt{T|A| \ln(\frac{|S||A|T}{\delta})} + CH^2|A||S|^2 \ln^2(\frac{|S||A|T}{\delta}) + \delta T^2
\end{aligned}$$

□

If we choose $\delta = \frac{1}{T^2}$, then the regret R_k is bounded by $\mathcal{O}(H^{1.5}|S|\sqrt{|A|T})$.

6 Reference

1. Azar, Mohammad Gheshlaghi, Ian Osband, and Rémi Munos. "Minimax regret bounds for reinforcement learning." In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 263-272. JMLR. org, 2017.